PROGRAM FOR ALLOWING COMPUTER TO CARRY OUT AUTOMATIC SELECTION OF TRANSLATION SYSTEM, AND COMPUTER-READABLE RECORDING MEDIUM RECORDING THE PROGRAM

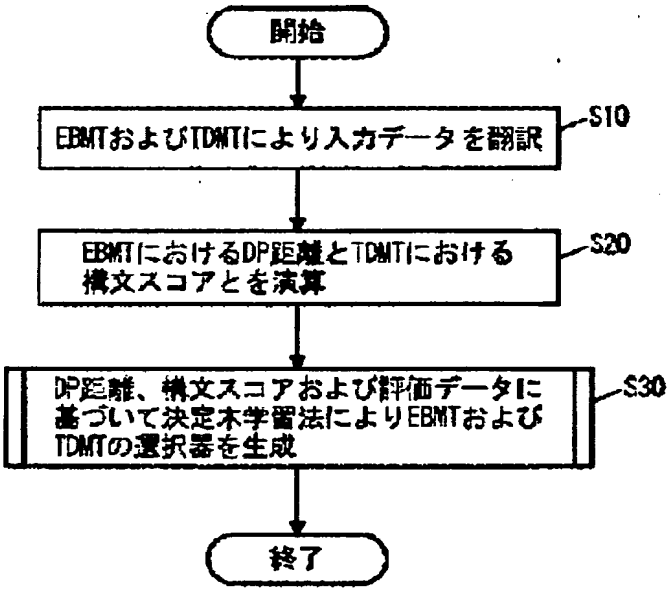
Patent number: JP2003263434
Publication date: 2003-09-19

Inventor: YASUDA YOSHIYUKI; SUGAYA FUMIAKI; TAKEZAWA TOSHIYUKI; YAMAMOTO SEIICHI
Applicant: ATR ADVANCED TELECOMM RES INST

Classification:
- international: *G06F17/28;* **G06F17/28**; (IPC1-7): G06F17/28
- european:
Application number: JP20020065365 20020311
Priority number(s): JP20020065365 20020311

Abstract of **JP2003263434**

PROBLEM TO BE SOLVED: To provide a program for allowing a computer to carry out automatic selection of a translation system suitable for the translation of input data, out of a plurality of translation systems.
SOLUTION: Input data is translated by two translation systems TDMT and EBMT (step S10). A sentence structure score showing similarity between the input data and examples in translating the input data by the EBMT, and a DP distance showing similarity between the input data and examples in translating the input data by the EBMT, are computed (step S20). Evaluation data showing whether the TDMT and EBMT are suitable for the translation of the input data, and the sentence structure score and DP distance computed in the step S20, are used to generate a selector for selecting the translation system suitable for the translation of the input data by a decision tree learning method (step S30).
COPYRIGHT: (C)2003,JPO



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－263434
（P2003－263434A）
(43)公開日　平成15年９月19日(2003.9.19)

(51)Int.Cl.'	識別記号	ＦＩ	テーマコード'（参考）
Ｇ０６Ｆ　17/28		Ｇ０６Ｆ　17/28	Ｚ　５Ｂ０９１

審査請求　未請求　請求項の数６　ＯＬ　（全 12 頁）

(21)出願番号　特願2002－65365(P2002－65365)

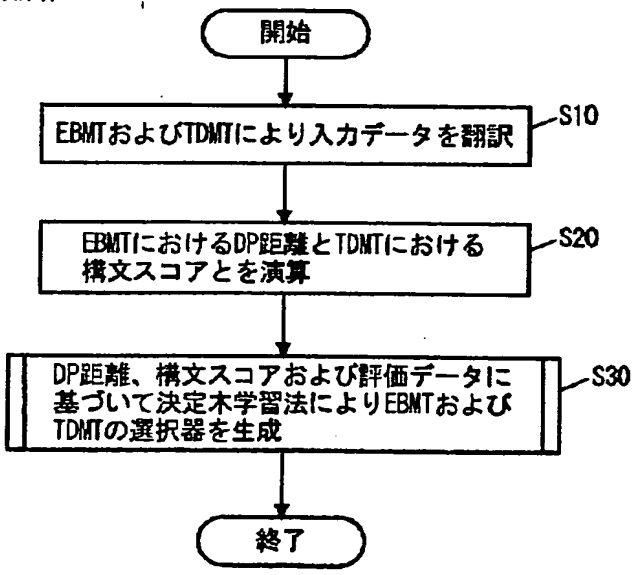
(22)出願日　平成14年３月11日(2002.3.11)

(71)出願人　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(72)発明者　安田　圭志
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　菅谷　史昭
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(74)代理人　100064746
弁理士　深見　久郎　（外４名）

最終頁に続く

(54)【発明の名称】　翻訳システムの自動選択をコンピュータに実行させるためのプログラム、およびそのプログラムを記録したコンピュータ読取り可能な記録媒体

(57)【要約】
【課題】　複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを提供する。
【解決手段】　２つの翻訳システムＴＤＭＴ，ＥＢＭＴによって入力データを翻訳する（ステップＳ１０）。そして、入力データをＴＤＭＴにより翻訳する際の入力データと用例との類似性を示す構文スコアと、入力データをＥＢＭＴにより翻訳する際の入力データと用例との類似性を示すＤＰ距離とを演算する（ステップＳ２０）。入力データの翻訳にＴＤＭＴ，ＥＢＭＴが適しているか否かを示す評価データと、ステップＳ２０において演算した構文スコアおよびＤＰ距離とを用いて決定木学習法により入力データの翻訳に適した翻訳システムを選択するための選択器を生成する（ステップＳ３０）。



1

【特許請求の範囲】
【請求項１】　ｍ（ｍは自然数）個の翻訳システムからｎ（ｎは自然数）個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムであって、
前記ｍ個の翻訳システムの各々により前記ｎ個の入力データの各々を翻訳する第１のステップと、
前記ｍ個の翻訳システムの各々が前記ｎ個の入力データの各々を翻訳する際の翻訳し易さを示すｎ個の指標を前記ｍ個の翻訳システムに対して演算する第２のステップと、
前記演算された第１番目のｎ個の指標乃至第ｍ番目のｎ個の指標と、前記ｎ個の入力データの翻訳に適した翻訳システムを前記ｎ個の入力データの各入力データごとに示すｎ個の評価データとに基づいて、決定木学習法により前記ｍ個の翻訳システムの各々に対して翻訳の適合性を演算する第３のステップとをコンピュータに実行させるためのプログラム。
【請求項２】　前記ｍ個の翻訳システムは、第１および第２の翻訳システムであり、
前記第１のステップは、
前記第１の翻訳システムを用いて、前記ｎ個の入力データの各々を構成する各単語を翻訳語に変換する第１のサブステップと、
前記第２の翻訳システムを用いて、前記ｎ個の入力データを構文構造によって分割した複数のパターンの各々を翻訳語に変換する第２のサブステップとを含み、
前記第２のステップは、
前記各単語を前記翻訳語に変換する際の前記各単語の意味と各単語に対応する翻訳語の意味との類似性を示すＤＰ距離を前記ｎ個の入力データの各々に対して演算する第３のサブステップと、
前記パターンを前記翻訳語に変換する際の前記パターンの意味と前記翻訳語の意味との類似性を示す構文スコアを前記ｎ個の入力データの各々に対して演算する第４のサブステップとを含む、請求項１に記載のコンピュータに実行させるためのプログラム。
【請求項３】　前記第１のステップにおいて前記第１および第２の翻訳システムにより翻訳されたそれぞれ第１および第２の翻訳データを外部へ出力する第４のステップをさらにコンピュータに実行させる、請求項２に記載のコンピュータに実行させるためのプログラム。
【請求項４】　前記ｎ個の評価データを外部から受付ける第５のステップをさらにコンピュータに実行させる、請求項３に記載のコンピュータに実行させるためのプログラム。
【請求項５】　前記評価データは、前記第１および第２の翻訳データに基づいて前記翻訳の正しさを人間により評価したデータである、請求項４に記載のコンピュータに実行させるためのプログラム。

2

【請求項６】　請求項１から請求項５のいずれか１項に記載のプログラムを記録したコンピュータ読取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、複数の翻訳システムから翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラム、およびそのプログラムを記録したコンピュータ読取り可能な記録媒体に関するものである。
【０００２】
【従来の技術】今までに、様々な機械翻訳システムが研究・開発されている。例えば、ＥＢＭＴ（Ｅｘａｍｐｌｅ　Ｂａｓｅｄ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ）、およびＴＤＭＴ（Ｔｒａｎｓｆｅｒ　Ｄｒｉｖｅｎ　Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ）が研究・開発されている。
【０００３】ＥＢＭＴは、１つの文に含まれる単語単位で変換すべき翻訳語を検索して入力された入力データを翻訳する翻訳システムである。また、ＴＤＭＴは、１つの文よりも短い、構文構造の基本単位である構成素境界パターンを単位として翻訳語に変換する翻訳システムである。
【０００４】
【発明が解決しようとする課題】しかし、これらの翻訳システムは、それぞれ、特定のドメインまたは特定の表現形式に適しているため、ある翻訳システムにより入力データを翻訳する翻訳器に、その翻訳システムによる翻訳に適さない入力データが入力されと、正確な翻訳ができないという問題が生じる。
【０００５】そこで、この発明は、かかる問題を解決するためになされたものであり、その目的は、複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを提供することである。
【０００６】また、この発明の別の目的は、複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを記録したコンピュータ読取り可能な記録媒体を提供することである。
【０００７】
【課題を解決するための手段および発明の効果】この発明によれば、ｍ（ｍは自然数）個の翻訳システムからｎ（ｎは自然数）個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、ｍ個の翻訳システムの各々によりｎ個の入力データの各々を翻訳する第１のステップと、ｍ個の翻訳システムの各々がｎ個の入力データの各々を翻訳する際の翻訳し易さを示すｎ個の指標をｍ個の翻訳システムに対して演算する第２のステップと、演算された第１番目

のｎ個の指標乃至第ｍ番目のｎ個の指標と、ｎ個の入力データの翻訳に適した翻訳システムをｎ個の入力データの各入力データごとに示すｎ個の評価データとに基づいて、決定木学習法によりｍ個の翻訳システムの各々に対して翻訳の適合性を演算する第３のステップとをコンピュータに実行させるためのプログラムである。

【０００８】好ましくは、ｍ個の翻訳システムは、第１および第２の翻訳システムであり、第１のステップは、第１の翻訳システムを用いて、ｎ個の入力データの各々を構成する各単語を翻訳語に変換する第１のサブステップと、第２の翻訳システムを用いて、ｎ個の入力データを構文構造によって分割した複数のパターンの各々を翻訳語に変換する第２のサブステップとを含み、第２のステップは、各単語を翻訳語に変換する際の各単語の意味と各単語に対応する翻訳語の意味との類似性を示すＤＰ距離をｎ個の入力データの各々に対して演算する第３のサブステップと、パターンを翻訳語に変換する際のパターンの意味と翻訳語の意味との類似性を示す構文スコアをｎ個の入力データの各々に対して演算する第４のサブステップとを含む。

【０００９】より好ましくは、ｍ（ｍは自然数）個の翻訳システムからｎ（ｎは自然数）個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、第１のステップにおいて第１および第２の翻訳システムにより翻訳されたそれぞれ第１および第２の翻訳データを外部へ出力する第４のステップをさらにコンピュータに実行させる。

【００１０】さらに好ましくは、ｍ（ｍは自然数）個の翻訳システムからｎ（ｎは自然数）個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、ｎ個の評価データを外部から受付ける第５のステップをさらにコンピュータに実行させる。

【００１１】さらに好ましくは、評価データは、第１および第２の翻訳データに基づいて翻訳の正しさを人間により評価したデータである。

【００１２】また、この発明によれば、プログラムを記録したコンピュータ読取り可能な記録媒体は、請求項１から請求項５のいずれか１項に記載のプログラムを記録したコンピュータ読取り可能な記録媒体である。

【００１３】従って、この発明によれば、複数の翻訳システムから入力データの翻訳に適した翻訳システムを選択することができる。

【００１４】

【発明の実施の形態】本発明の実施の形態について図面を参照しながら詳細に説明する。なお、図中同一または相当部分には同一符号を付してその説明は繰返さない。

【００１５】この発明によるプログラムは、複数の翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する。そして、この発明の実施の形態においては、２つの翻訳システムから入力データの翻訳に適した翻訳システムを自動選択するプログラムについて説明し、２つの翻訳システムは、ＴＤＭＴおよびＥＢＭＴである。

【００１６】図１は、２つの翻訳システム（ＴＤＭＴおよびＥＢＭＴ）の概要を説明するための図である。ＥＢＭＴでは、多くの用例を含むコーパスから、各入力データ（Ｄ１，Ｄ２，・・・，Ｄｎ）に最も類似した用例を検索し、最も類似した用例を翻訳に用いる。コーパスは、多くの原言語用例（入力側の言語）と目的言語用例（出力側言語）のペアから成る。最も類似した用例は、各入力とコーパス内の各原言語用例とのＤＰ距離に基づいて検索される。そして、ＤＰ距離は次式で定義される。

【００１７】

【数１】

$$P_{DP} = \frac{I + D + 2\Sigma D_{semantic}}{L_{input} + L_{example}} \quad \cdots (1)$$

【００１８】但し、L_{input} は入力を表わし、$L_{example}$ は、コーパス内の用例の単語数を表わし、Ｉは、入力文とコーパス内の用例とをＤＰマッチングで比較した時の挿入語数を表わし、Ｄは、入力文とコーパス内の用例とをＤＰマッチングで比較した時の脱落語数を表わし、$D_{semantic}$ は、単語間の意味距離である。

【００１９】式（１）により定義されるＤＰ距離P_{DP}は、入力文と用例とがどの程度類似しているかを示し、「０」から「１」までの値をとる。そして、入力文と用例とが完全に一致した場合、ＤＰ距離P_{DP}は「０」となる。

【００２０】従って、入力文Ｄ１～Ｄｎの各々について、最も類似した用例の各々をコーパスから抽出するとき、コーパスに含まれる用例の全てについて式（１）を用いてＤＰ距離P_{DP}を演算し、ＤＰ距離P_{DP}が最も小さい用例をコーパスから抽出する。

【００２１】翻訳の手順を以下に述べる。ここで、ある入力データＤｉ（１≦ｉ≦ｎ）とし、Ｄｉは単語Ｗ１，・・・，Ｗｐ（ｐは自然数）の系列から成るとする。コーパスの検索により得られた原言語用例をＧとし、Ｇは単語ＷＧ１，・・・，ＷＧｐの系列からなるとする。この原言語文に対応する目的言語用例をＭとし、Ｍは単語ＷＭ１，・・・，ＷＭｐの系列から成るとする。入力データＤｉと、原言語用例Ｇとが完全に一致した場合は、原言語用例に対応する目的言語用例Ｍをそのまま出力する。入力データＤｉと、原言語用例Ｇとが完全に一致しない場合は、まず、入力データＤｉと原言語用例Ｇとを比較して異なる単語を検出する。ここで、入力データＤｉの単語Ｗｊ（１≦ｊ≦ｐ）と目的言語用例Ｇの単語ＷＧｊのみが異なり、それ以外の単語については同じであるとすると、ＷＧｊに対応する単語を辞書などの情報を用いてＷｊに対応する目的言語の単語ＷＭ＿ｎｅｗｋを

求め、目的言語用例Ｍにおいて単語WM_kを単語WM_new_kに置き換えて出力する。

【００２２】ＴＤＭＴは、入力データＤの各データＤ１，Ｄ２，・・・，Ｄｎよりも短く、かつ、構文構造の基本単位である構成素境界パターン（以下、「パターン」と言う。）を単位としてコーパスから変換知識を学習し、その学習した変換知識を用いて入力データＤを翻訳する翻訳システムである。すなわち、入力データＤ（Ｄ１，Ｄ２，・・・，Ｄｎ）は、ｑ（ｑは自然数）個のパターンＰＴ１～ＰＴｑに分割され、パターンＰＴ１～ＰＴｑの各々を、コーパスから選択したパターンＴＰＴ１～ＴＰＴｑにそれぞれ変換することにより、入力データＤは翻訳される。

【００２３】より具体的には、パターンＰＴ１～ＰＴｑの各々をパターンＴＰＴ１～ＴＰＴｑに変換するとき、次式により定義される構文スコアを演算する。

【００２４】

【数２】

$$P_{TDMT} = \underset{(Pj)\in P}{\operatorname{argmin}} \sum_i S\cdot(bi, Pj) \quad \cdots (2)$$

【００２５】但し、ＰはＴＤＭＴが持つパターンの集合を表わし、b_i（ｉは自然数）は入力文を解析して得られる部分文を表わし、Ｓはパターンと部分文との意味距離を表わす。

【００２６】構文スコアP_{TDMT}は、「０」以上の値を採り、値が小さい方が意味の近いパターンであることを示す。

【００２７】従って、パターンＰＴ１～ＰＴｑの各々について、コーパスに含まれる全ての用例について式（２）を用いて構文スコアP_{TDMT}を演算し、構文スコアP_{TDMT}が最小となるパターンがコーパスから選択される。

【００２８】このように、ＴＤＭＴは、入力データＤの各データＤ１，Ｄ２，・・・，Ｄｎよりも短いパターンを基本単位として入力データＤを翻訳する翻訳システムである。

【００２９】なお、ＴＤＭＴでは、複数の文からなる長い発話や、話し言葉に見られる文法法規から逸脱した入力等の原因で、変換を適用する部分文の組合せにより１つの構文木を作れなかった場合は、最も整合性のとれた部分に分割し、部分毎に翻訳を行なう。

【００３０】図２は、この発明によるプログラムが、上述した２つの翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する際の概念を示す概念図である。

【００３１】まず、学習文１０がＴＤＭＴ翻訳器２０およびＥＢＭＴ翻訳器３０に入力される。ＴＤＭＴ翻訳器２０は、学習文１０を上述したＴＤＭＴにより翻訳し、翻訳文であるＴＤＭＴ出力と、構文スコアP_{TDMT}とを出力する。また、ＥＢＭＴ翻訳器３０は、学習分１０を上述したＥＢＭＴにより翻訳し、翻訳文であるＥＢＭＴ出力と、ＤＰ距離P_{DP}とを出力する。

【００３２】評価器４０は、ＴＤＭＴ翻訳器２０からのＴＤＭＴ出力と、ＥＢＭＴ翻訳器３０からのＥＢＭＴ出力とを受け、ＴＤＭＴ出力およびＥＢＭＴ出力を、翻訳一対比較法により評価する。つまり、人間が学習文１０の翻訳文としてＴＤＭＴ出力とＥＢＭＴ出力とのうち、どちらが適しているかを翻訳前の学習文１０と翻訳文であるＴＤＭＴ出力（またはＥＢＭＴ出力）とを一対一に比較して評価する。そして、評価器４０は、評価結果を出力する。

【００３３】決定木学習器５０は、ＴＤＭＴ翻訳器２０から構文スコアP_{TDMT}を受け、ＥＢＭＴ翻訳器３０からＤＰ距離P_{DP}を受け、評価器４０から評価結果を受ける。そして、決定木学習器５０は、構文スコアP_{TDMT}、ＤＰ距離P_{DP}、および評価結果に基づいて、決定木学習法を用いて選択器を生成する。

【００３４】自動選択器６０は、決定木学習器５０から選択器に基づいて、学習文１０の翻訳に適した翻訳システムを選択する。

【００３５】このように、この発明によるプログラムは、ＴＤＭＴにおいて変換すべきパターンの見つけ易さ、すなわち、翻訳し易さを示す構文スコアP_{TDMT}、ＥＢＭＴにおいて変換すべき単語の見つけ易さ、すなわち、翻訳のし易さを示すＤＰ距離P_{DP}、および評価結果に基づいて、適した翻訳システムを選択するための選択器を生成することを特徴とする。

【００３６】図３は、この発明によるプログラムがコンピュータに実行させるためのステップを説明するためのフローチャートである。動作が開始されると、ＥＢＭＴおよびＴＤＭＴにより入力データを翻訳する（ステップＳ１０）。そして、ＥＢＭＴにおけるＤＰ距離P_{DP}と、ＴＤＭＴにおける構文スコアP_{TDMT}とを演算し（ステップＳ２０）、構文スコアP_{TDMT}、ＤＰ距離P_{DP}、および評価結果に基づいて、決定木学習法を用いて選択器を生成する（ステップＳ３０）。これにより、一連の動作は終了する。

【００３７】図４は、図３に示すフローチャートのステップＳ３０における決定木学習法による選択器の生成の詳細な動作を説明するためのフローチャートである。図４に示すフローチャートについて説明する前に、前提について説明する。学習データＬは、ＴＤＭＴにおける構文スコアP_{TDMT}であるＴと、ＥＢＭＴにおけるＤＰ距離P_{DP}であるＥと、評価データであるＣとから成る。すなわち、Ｌ＝（Ｔ，Ｅ，Ｃ）である。

【００３８】また、構文スコアＴ、ＤＰ距離Ｅ、および評価データＣは、それぞれ、ｎ個の要素から成る。すなわち、$T = |t_1, t_2, \cdots t_n|$、$E = |e_1, e_2, \cdots e_n|$、$C = |c_1, c_2, \cdots c_n|$である。そして、$T = |t_1, t_2, \cdots t_n|$、$E =$

$|e_1, e_2, \cdots e_n|$ 、およびC＝ $|c_1, c_2, \cdots c_n|$ の各要素は相互に対応関係を有する。例えば、t_1、e_1、およびc_1は相互に対応関係を有する。その他の要素についても同様である。

【００３９】さらに、評価データCの要素$c_1, c_2, \cdots c_n$は、「０」か「１」のいずれかの値を採る。そして、「０」はＴＤＭＴの翻訳システムが優勢であることを示し、「１」はＥＢＭＴの翻訳システムが優勢であることを示す。

【００４０】なお、実際には、２つの翻訳システムＴＤＭＴ，ＥＢＭＴを比較した場合、同等と評価される場合もあるが、このようなデータは学習データＬに用いられない。

【００４１】図３に示すステップＳ２０の後、学習データＬ＝（Ｔ，Ｅ，Ｃ）が渡される（ステップＳ３１）。そして、構文スコアＴの全てのメンバー$t_1, t_2, \cdots t_n$の値を閾値の候補として、構文スコアＴに対する利得比を求める。この利得比の求め方について後述する。

【００４２】利得比の演算が終了すると、全てのメンバー$t_1, t_2, \cdots t_n$を閾値とした場合の構文スコアＴに対する利得比の中から最大となる利得比T_{gain}を抽出し、T_{gain}を与える閾値T_{hT}を抽出する。

【００４３】ＤＰ距離Ｅについても、構文スコアＴの場合と同様に最大となる利得比E_{gain}と、E_{gain}を与える閾値T_{hE}とを求める。そして、T_{gain}をE_{gain}と比較し、大きい方を最大利得比MAX_{gain}とする（ステップＳ３２）。

【００４４】ここで、利得比の演算について説明する。例としてＴに対する利得比を、メンバーt_kを閾値として求める場合について説明する。Ｔの値により、Ｔと、それに対応するＣをソートし、t_kを境界としてＣ＝$|c_1, c_2, \cdots c_n|$を２つに分割する。分割後のＣをそれぞれ$C_{div(1)}$とと$C_{div(2)}$とで表わす（$C=C_{div(1)}+C_{div(2)}$）。

【００４５】ここで、値が「０」となるＣのメンバーの数をｆｒｅｑ（C_0，Ｃ）で表わし、値が「１」となるＣのメンバーの数をｆｒｅｑ（C_1，Ｃ）で表わす。また、$|C|$をＣに含まれるメンバーの数とすると、Ｃの平均情報量は次式により求められる。

【００４６】

【数３】

$$info(c) = -\sum_{j=0}^{1} \frac{freq(Cj, C)}{|C|} \times \log_2\left(\frac{freq(Cj, C)}{|C|}\right) \quad \cdots (3)$$

【００４７】ｉｎｆｏ（Ｃ）は、Ｃ内にある１つの事例に属するクラスを道程するのに必要な情報量の平均値となる。

【００４８】また、分割後のデータ$C_{div(1)}$、$C_{div(2)}$に同様な評価を考えると、部分集合上で荷重平均をとって、次式により求められる。

【００４９】

【数４】

$$infot_k(c) = \sum_{j=0}^{1} \frac{|C_{div(j)}|}{|C|} \times info(C_{div(j)}) \quad \cdots (4)$$

【００５０】さらに、分割情報量は次式により求められる。

【００５１】

【数５】

$$Split\_info(t_k) = -\sum_{j=0}^{1} \frac{|C_{div(j)}|}{|C|} \times \log_2 \frac{|C_{div(j)}|}{|C|} \quad \cdots (5)$$

【００５２】そして、利得比基準は、これまでの式を用いて次式により求められる。

【００５３】

【数６】

$$gain\_ratio(t_k) = \frac{info(c) - infot_k(c)}{Split\_info(t_k)} \quad \cdots (6)$$

【００５４】従って、ステップＳ３２においては、上述した式（３）～（６）を用いて構文スコアＴおよびＤＰ距離Ｅに対して利得比が演算され、最終的に、最大利得比MAX_{gain}が演算される。

【００５５】ステップＳ３２の後、最大利得比MAX_{gain}が「０」以下、またはＣのメンバー$c_1, c_2, \cdots c_n$の値が全て同じか否かを判定する（ステップＳ３３）。そして、最大利得比MAX_{gain}が「０」以下である場合、またはＣのメンバー$c_1, c_2, \cdots c_n$の値が全て同じである場合、決定木学習法により選択器を生成する動作は終了する（ステップＳ３４）。

【００５６】ステップＳ３３において、最大利得比MAX_{gain}が「０」よりも大きく、かつ、Ｃのメンバー$c_1, c_2, \cdots c_n$の値が相互に不一致である場合、学習データＬは２つの学習データＬ０，Ｌ１に分割される（ステップＳ３５）。この場合、ステップＳ３２における最大利得比$MAX_{gain}=T_{gain}$であるなら、閾値T_{hT}を用いて学習データＬを学習データＬ０，Ｌ１に分割し、ステップＳ３２における最大利得比$MAX_{gain}=E_{gain}$であるなら、閾値T_{hE}を用いて学習データＬを学習データＬ０，Ｌ１に分割する。

【００５７】この分割の具体例を図５に示す。閾値$T_{hT}=t_k$であるときＴ＝$|t_1, t_2, \cdots, t_k, \cdots t_n|$を閾値$t_k$により２つのデータＴ０＝$|t_1, t_2, \cdots t_k|$，Ｔ１＝$|t_{k+1}, t_2, \cdots t_n|$に分割する。また、Ｅ＝$|e_1, e_2, \cdots, e_k, \cdots e_n|$を閾値$t_k$に対応するメンバー$e_k$を基準にして２つのデータＥ０＝$|e_1, e_2, \cdots e_k|$，Ｅ１＝$|e_{k+1}, e_2, \cdots e_n|$に分割する。さらに、Ｃ＝$|c_1, c_2, \cdots, c_k, \cdots c_n|$についても、同様に、２つのデータＣ０＝$|c_1, c_2, \cdots c_k|$，Ｃ１＝$|c_{k+1}, c_2, \cdots c_n|$に分割する。

【００５８】従って、学習データＬ０＝（$|t_1, t_2, \cdots t_k|$，$|e_1, e_2, \cdots e_k|$，$|c_1, c_2,$

・・・ck|）であり、学習データＬ１＝（|tk+1，t2，・・・tn|，|ek+1，e2，・・・en|，|ck+1，c2，・・・cn|）である。

【００５９】閾値ThEを用いて学習データＬを分割する場合も同様である。図４に示すフローチャートのステップＳ３５において、学習データＬが学習データＬ０，Ｌ１に分割されると、学習データＬ０，Ｌ１の各々についてステップＳ３１～Ｓ３５が繰返し行なわれる（ステップＳ４０，Ｓ５０）。

【００６０】ステップＳ４０，Ｓ５０の各々に含まれるステップＳ３３において、最大利得比ＭＡＸgain が「０」よりも大きく、かつ、Ｃのメンバーc1，c2，・・・cnの値が相互に不一致である場合、学習データＬ０，Ｌ１の各々は、２つの学習データに分割される。そして、分割された各学習データを対象としてステップＳ３１～Ｓ３５が繰返し行なわれる。

【００６１】すなわち、学習データＬを最初の学習データとしていずれの翻訳システムが適しているかの判定を開始し、ツリー状に学習データＬを分割し、その分割した学習データの各々を対象としていずれの翻訳システムが適しているかの判定を行なう。

【００６２】そして、各学習データを対象とした判定のステップＳ３３において、最大利得比ＭＡＸgain が「０」以下である場合、またはＣのメンバーc1，c2，・・・cnの値が全て同じである場合に決定木学習法により選択器を生成する動作は終了する。

【００６３】選択器を生成することは、構文スコア（「ＴＤＭＴ距離」とも言う。）とＤＰ距離（「ＥＢＭＴ距離」とも言う。）とによって形成される二次元空間において、ＴＤＭＴに適した領域および／またはＥＢＭＴに適した領域に分割することに相当する。

【００６４】例えば、図６に示すように、点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域を、点Ｅ－Ｆ－Ｇ－Ｈ－Ｉ－Ｊ－Ｄ－Ｃによって囲まれる領域と、点Ａ－Ｂ－Ｅ－Ｆ－Ｇ－Ｈ－Ｉ－Ｊによって囲まれる領域とに分割する。

【００６５】図４に示すフローチャートのステップＳ３３において、Ｃのメンバーc1，c2，・・・cnの値が全て同じであるとき、ＴＤＭＴおよびＥＢＭＴのいずれか一方が入力データの翻訳に適した翻訳システムであることを示すので、点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域は分割されずに、決定木学習法による選択器の生成は終了する。

【００６６】そして、ステップＳ３５において、閾値ThEによって学習データＬを学習データＬ０，Ｌ１に分割することは、図６に示す点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域において、ＥＢＭＴ距離を示す横軸上の閾値ThEの値に相当する点Ｋから縦方向に線１００を引き、点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域を、点Ｃ－Ｄ－Ｊ－Ｋによって囲まれる領域と、点Ａ－Ｂ－Ｋ－Ｊによって囲まれる領域とに分割することに相当する。

【００６７】また、ステップＳ３５において、最大利得比ＭＡＸgain を与えるのがＥgain であり、閾値ThEによって学習データＬを学習データＬ０，Ｌ１に分割することは、図６に示す点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域において、ＴＤＭＴ距離を示す縦軸上の閾値ThEの値に相当する点から横方向に線を引き、点Ａ－Ｂ－Ｃ－Ｄによって囲まれる領域を２つの領域に分割することに相当する。

【００６８】ステップＳ４０，Ｓ５０の各々において、学習データＬ０，Ｌ１の各々がさらに２つの学習データＬ０１，Ｌ０２；Ｌ１１，Ｌ１２に分割されると、上述した方法と同じ方法によって線１０１，１０２を順次引く。そして、学習データＬ０１，Ｌ０２および学習データＬ１１，Ｌ１２のいずれかの学習データＬ０１，Ｌ０２（または学習データＬ１１，Ｌ１２）において、さらに、学習データの分割が行なわれると、線１０３，１０４を順次引く。その結果、点Ｈ－Ｍ－Ｋ－Ｉによって囲まれる領域および点Ｆ－Ｅ－Ｍ－Ｇによって囲まれる領域が生成される。

【００６９】この場合、分割された学習データにおいて、評価データのメンバーの値が全て同じ値になるか、演算された最大利得比ＭＡＸgain が「０」以下になると、その分割された学習データの翻訳に適した翻訳システムの決定は終了する。最大利得比ＭＡＸgain が「０」以下であるとき、対象とする学習データの翻訳に適した翻訳システムを決定する動作が終了することとしているのは、最大利得比ＭＡＸgain が「０」以下になると、対象とする学習データを構成するＴＤＭＴ距離ＴとＥＢＭＴ距離Ｅとによって決定される領域をさらに分割することができないからである。

【００７０】このように、対象とする学習データを構成する評価データのメンバーの値が全て同じ値になるか、対象とする学習データを構成するＴＤＭＴ距離ＴとＥＢＭＴ距離Ｅとによって決定される領域をさらに分割することができなくなるまで、学習データＬをツリー状に次々と分割することによって、ＴＤＭＴによる翻訳に適した領域とＥＢＭＴによる翻訳に適した領域とに分割することができる。

【００７１】図６に示す例では、点Ｅ－Ｆ－Ｇ－Ｈ－Ｉ－Ｊ－Ｄ－Ｃによって囲まれる領域は、ＥＢＭＴによる翻訳に適した領域であり、点Ａ－Ｂ－Ｅ－Ｆ－Ｇ－Ｈ－Ｉ－Ｊによって囲まれる領域は、ＴＤＭＴによる翻訳に適した領域である。

【００７２】従って、自動選択器６０は、決定木学習法により図６に示すような各翻訳システムに適した領域に分割されたデータ（「選択器」と言う。）を受取ることによって入力データに適した翻訳システムを用いて入力データを翻訳できる。

【００７３】決定木学習法による翻訳システムの選択を実際に行なった例について説明する。表１は、学習セッ

トおよび評価セットについて、各入力発話に対するＴＤＭＴおよびＥＢＭＴの出力を、評価者が一対比較した結果を示す。

【００７４】

【表１】

	学習セット	評価セット
EBMT 勝	189	189
TDMT 勝	175	186
同等	144	135

【００７５】表１に示すように、比較結果は、ＥＢＭＴ優位（表１中の「ＥＢＭＴ勝」）、ＴＤＭＴ優位（表１中の「ＴＤＭＴ勝」）および同等によって表わされる。

また、表１中の数字は実際の文数を表わしている。元々の学習セットは５０８文、評価セットは５１０文からなるが、「同等」と判断されたデータについては、自動選択を行なう場合に、どちらが選択されても翻訳品質に影響がでないため学習および評価には用いていない。実際の学習および評価に用いたのは、一対比較により優劣が決定されたデータ（表１中の下線部）であり、学習セットは３６４文、評価セットは３７５文である。

【００７６】表２は、決定木学習データの構造を示す。ＥＢＭＴ距離、ＴＤＭＴ距離、および人間による評価結果が示されている。

【００７７】

【表２】

Input		Teacher
EBMT (DP Distance)	TDMT (Syntax Score)	Result of Paired Comparison Method
1	4.57	TDMT
0	0	TDMT
0.2	25	EBMT
0.2	0.67	TDMT
0	0.83	EBMT
:	:	:
0	1.33	EBMT

【００７８】また、表３は、学習された決定木により自動選択を行なった結果を示す。

【００７９】

【表３】

		人間による選択	
		EBMT	TDMT
Decision Tree	EBMT	135(36%)	36(9.6%)
	TDMT	54(14.4%)	150(40%)

【００８０】表３中の数字は、評価文の数を表わしており、括弧内の数字は評価セット全体（３７５文）に対する割合を表わしている。また、表３の下線部が正しく選択された結果を表わしており、正しく選択される割合は７６％となっている。

【００８１】図７は、表１～表３に示した場合を図示したものである。横軸はＥＢＭＴ距離であり、縦軸はＴＤＭＴ距離である。なお、直線１０５は、従来の方法によってＴＤＭＴによる翻訳に適した領域とＥＢＭＴによる翻訳に適した領域とに分割する場合の境界線である。また、「●」は人間が入力データの翻訳に適している翻訳システムとしてＴＤＭＴを選択したことを表わし、「▲」は人間が入力データの翻訳に適している翻訳システムとしてＥＢＭＴを選択したことを表わす。さらに、領域１１０～１１２がＴＤＭＴによる翻訳に適した領域であり、それ以外の領域１１３はＥＢＭＴによる翻訳に適した領域である。

【００８２】ＴＤＭＴのプロットは、殆ど、領域１１０～１１２に含まれており、上述した決定木学習法は人間が判断した場合とよい一致を示す。但し、ＥＢＭＴのプロットも、領域１１２に含まれており、決定木学習法による翻訳システムの選択には、若干の誤りがあることがわかる。

【００８３】全体として、決定木学習法による翻訳システムの選択が人間による評価結果と一致する割合は上述したように７６％である。

【００８４】従来は、ＴＤＭＴによる翻訳に適した領域および／またはＥＢＭＴによる翻訳に適した領域への分割は直線１０５によってのみ可能であり、２つの領域の境界は直線的に決定されていたが、決定木学習法を用いることにより、２つの領域の境界は階段状に決定できるようになった。これにより、各種の入力データに適した翻訳システムを選択可能になった。

【００８５】図８は、ＥＢＭＴ単体、ＴＤＭＴ単体、ＥＢＭＴとＴＤＭＴとを決定木学習法により選択した場合、および人間がＥＢＭＴとＴＤＭＴとを選択した場合の翻訳性能を、ＴＯＥＩＣスコア６８５のＴＯＥＩＣ受験者により評価した結果を示す。システムによる翻訳がＴＯＥＩＣ受験者による翻訳よりも優れている場合（図８中の「ＭＴ勝」）、システムによる翻訳とＴＯＥＩＣ受験者による翻訳が同等の場合（図８中の「Ｅｖｅｎ」）、およびＴＯＥＩＣ受験者による翻訳が翻訳シス

テムによる翻訳よりも優れている場合（図８中の「Ｈｕｍａｎ勝」）の３種類の結果が得られる。

【００８６】図８の横軸において、左からＥＢＭＴ単体での評価結果、ＴＤＭＴ単体での評価結果、決定木学習法により選択を行なった場合の評価結果、および人間が選択を行なった場合（すなわち、理想的な翻訳システムの選択を実現できた場合）の評価結果を示す。

【００８７】図８においては、ＥＢＭＴ単体およびＴＤＭＴ単体は、それぞれ、評価結果の割合がほぼ同じであり、ほぼ同等の翻訳性能である。決定木学習法により翻訳システムを選択した場合は、人間により翻訳システムを選択した場合よりも劣るが、ＥＢＭＴ単体またはＴＤＭＴ単体の場合と比較して、翻訳システムによる翻訳が優位となる文の割合が増加し、ＴＯＥＩＣ受験者による翻訳が優位となる文の割合が減少している。従って、決定木学習法により翻訳システムを選択し、ＥＢＭＴとＴＤＭＴとを併用することにより、ＥＢＭＴまたはＴＤＭＴを単体で用いた場合よりも翻訳性能が改善される。

【００８８】上記においては、ＴＤＭＴおよびＥＢＭＴという２つの翻訳システムから入力データの翻訳に適した翻訳システムを決定木学習法により選択する場合について説明したが、この発明においては、一般にｍ（ｍは自然数）個の翻訳システムから入力データの翻訳に適した翻訳システムを決定木学習法により選択してもよい。

【００８９】図９は、ｍ個の翻訳システム１～ｍから決定木学習法により選択された翻訳システムを用いて入力データを翻訳する場合の概念図である。自動選択器６０は、決定木学習法により生成された選択器に基づいて、翻訳システム１～ｍから入力データの翻訳に適した翻訳システムを選択し、その選択した翻訳システムにより入力データを翻訳して出力データを出力する。

【００９０】このように、入力データを翻訳する際の翻訳システムの数が増加することにより、入力データの翻訳に適した翻訳システムの組合わせは増加し、入力データをより正確に翻訳できる。

【００９１】この発明によるプログラムは、図１０に示すパーソナルコンピュータで実行される。図１０は、パーソナルコンピュータの概略ブロック図である。パーソナルコンピュータ９０は、データバスＢＳと、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）９１と、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）９２と、ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）９３と、シリアルインタフェース９４と、端子９５と、ＣＤ－ＲＯＭドライブ９６と、ディスプレイ９７と、キーボード９８とを備える。

【００９２】ＣＰＵ９１は、ＲＯＭ９３に格納されたプログラムをデータバスＢＳを介して読出す。また、ＣＰＵ９１は、シリアルインタフェース９４、端子９５およびインターネット網を介して取得したプログラム、またはＣＤ（Ｃｏｍｐａｃｔ　Ｄｉｓｋ）９９からＣＤ－ＲＯＭドライブ９６を介して読出したプログラムをＲＯＭ９３に格納する。さらに、ＣＰＵ９１は、キーボード９８から入力されたユーザからの指示を受付ける。

【００９３】ＲＡＭ９２は、ＣＰＵ９１が各種の処理を行なう際のワークメモリである。ＲＯＭ９３は、プログラムおよびコーパス等を格納する。シリアルインタフェース９４は、データバスＢＳと端子９５との間でデータのやり取りを行なう。

【００９４】端子９５は、ケーブルによってパーソナルコンピュータ９０をモデム（図示せず）に接続するための端子である。ＣＤ－ＲＯＭドライブ９６は、ＣＤ９９に記録されたプログラムを読出す。ディスプレイ９７は、各種の情報を視覚情報としてユーザに与える。キーボード９８は、ユーザからの指示を入力する。

【００９５】ＣＰＵ９１は、キーボード９８から入力されたユーザのプログラム実行要求に応じて、データバスＢＳを介してＲＯＭ９３からプログラムを読出し、その読出したプログラムを実行する。このプログラムの実行により入力データはＴＤＭＴまたはＥＢＭＴによって翻訳され、その翻訳結果はディスプレイ９７に表示される。これにより、人間はディスプレイ９７を見て入力データの翻訳に適した翻訳システム（ＴＤＭＴおよびＥＢＭＴのいずれか）を選択する。

【００９６】また、ＣＰＵ９１は、ＲＡＭ９２を用いてＥＢＭＴ距離およびＴＤＭＴ距離を演算し、その演算結果をＲＯＭ９３に格納する。そして、ＣＰＵ９１は、キーボード９８を介して評価データが入力されると、ＲＯＭ９３に格納されたＴＤＭＴ距離およびＥＢＭＴ距離をデータバスＢＳを介して読出し、上述した決定木学習法により翻訳システムの選択を行なう。翻訳システムの選択結果はディスプレイ９７に表示される。

【００９７】今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は、上記した実施の形態の説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【図面の簡単な説明】

【図１】　２つの翻訳システム（ＴＤＭＴおよびＥＢＭＴ）の概要を説明するための図である。

【図２】　複数の翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する際の概念を示す概念図である。

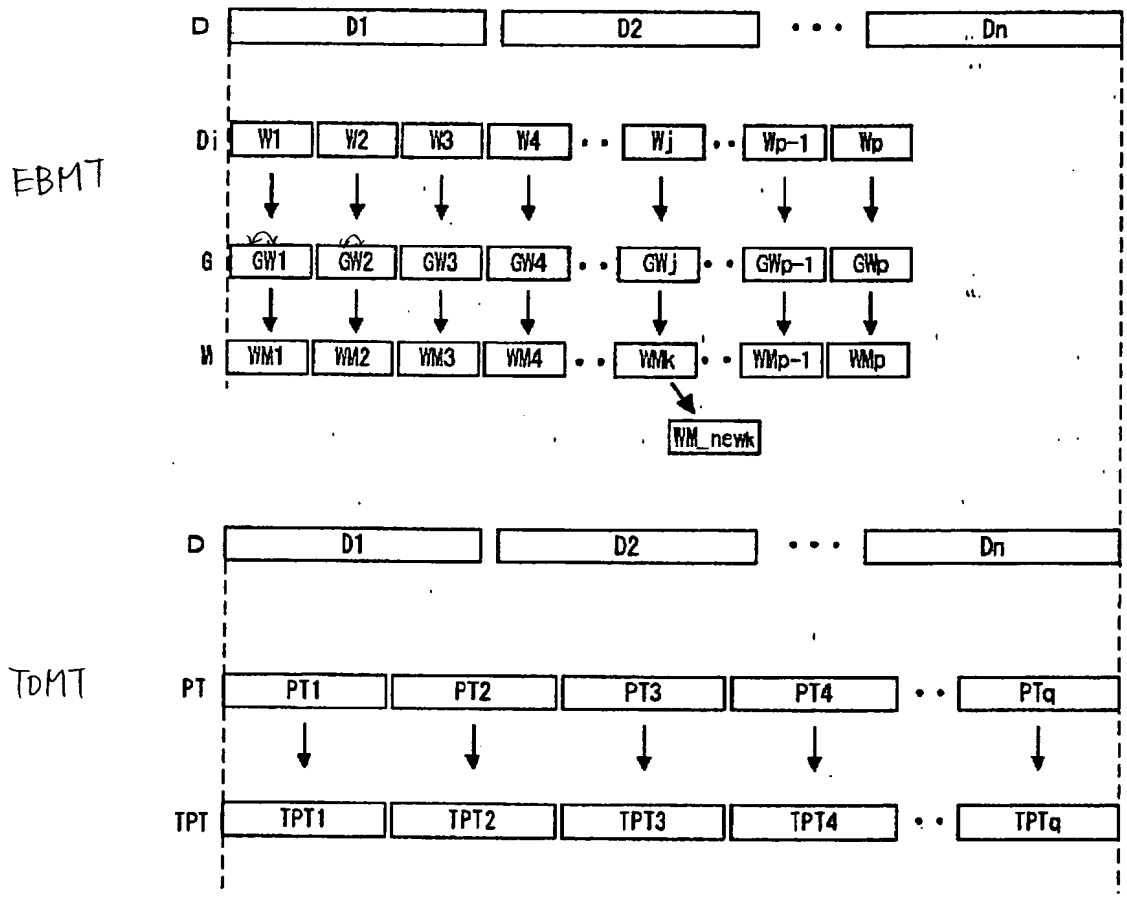
【図３】　この発明によるプログラムがコンピュータに実行させるためのステップを説明するためのフローチャートである。

【図４】　図３に示すフローチャートのステップＳ３０の詳細な動作を説明するためのフローチャートである。

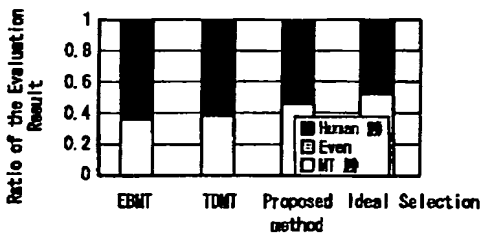
【図５】　学習データの分割を説明するための図である。

【図６】　入力データの翻訳に適した翻訳システムの領域を示す図である。
【図７】　入力データの翻訳に適した翻訳システムの領域の例を示す図である。
【図８】　決定木学習法による翻訳システムの評価結果を示す図である。
【図９】　複数の翻訳システムを用いた翻訳の概念図である。
【図１０】　パーソナルコンピュータの概略ブロック図である。
【符号の説明】
１０　学習文、２０　ＴＤＭＴ翻訳器、３０　ＥＢＭＴ翻訳器、４０　評価器、５０　決定木学習器、６０　自動選択器、９０　パーソナルコンピュータ、９１　ＣＰＵ、９２　ＲＡＭ、９３　ＲＯＭ、９４　シリアルインタフェース、９５　端子、９６　ＣＤ－ＲＯＭドライブ、９７　ディスプレイ、９８　キーボード、９９　ＣＤ、１００～１０６　線、１１０～１１２　領域。

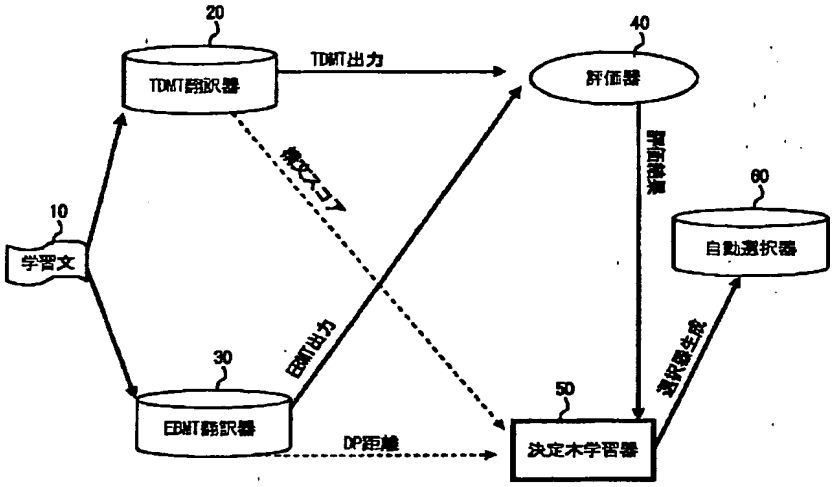
【図１】



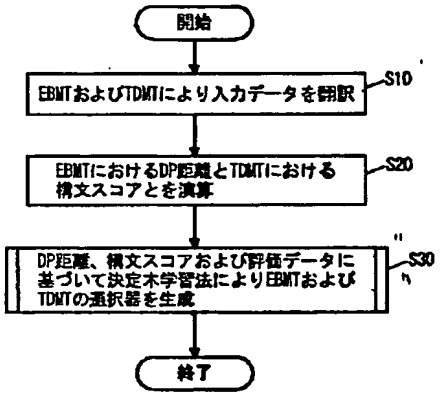
【図８】



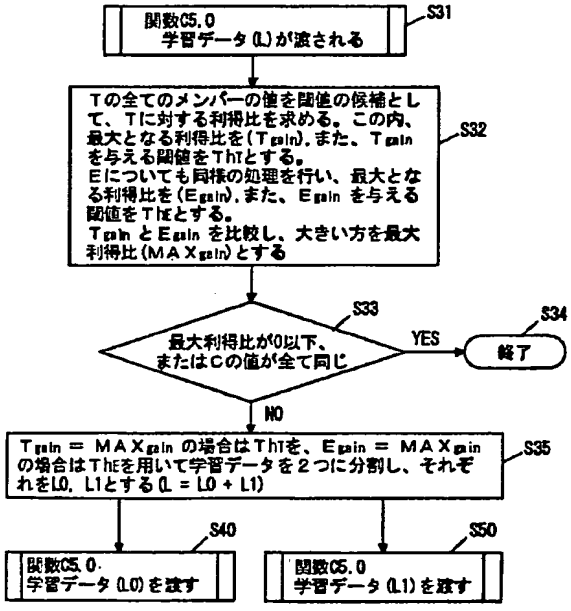
【図２】


20
TDMT翻訳器
TDMT出力
40
評価器
学習文
10
構文スコア
学習結果
60
自動選択器
EBMT出力
30
EBMT翻訳器
50
決定木学習器
選択器生成
DP距離

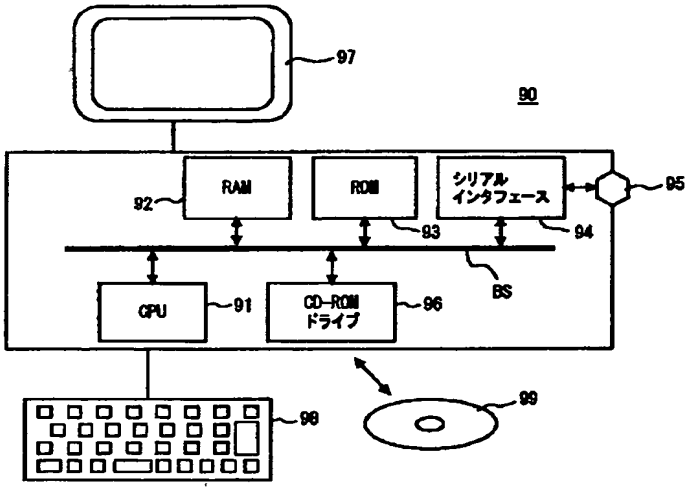
【図３】


開始
EBMTおよびTDMTにより入力データを翻訳
S10
EBMTにおけるDP距離とTDMTにおける構文スコアとを演算
S20
DP距離、構文スコアおよび評価データに基づいて決定木学習法によりEBMTおよびTDMTの選択器を生成
S30
終了

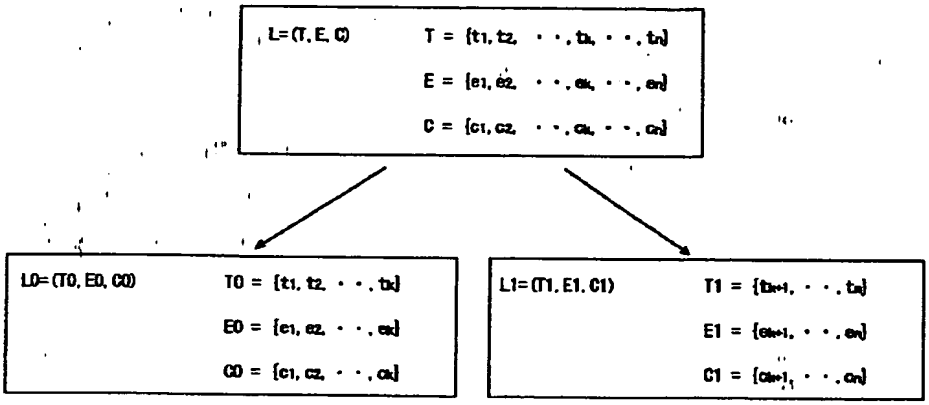
【図４】


関数C5.0
学習データ(L)が渡される
S31
Tの全てのメンバーの値を閾値の候補として、Tに対する利得比を求める。この内、最大となる利得比を(Tgain)、また、Tgainを与える閾値をThTとする。
Eについても同様の処理を行い、最大となる利得比を(Egain)、また、Egainを与える閾値をThEとする。
TgainとEgainを比較し、大きい方を最大利得比(MAXgain)とする
S32
最大利得比が0以下、またはCの値が全て同じ
S33
YES
S34
終了
NO
Tgain = MAXgainの場合はThTを、Egain = MAXgainの場合はThEを用いて学習データを２つに分割し、それぞれをL0、L1とする(L = L0 + L1)
S35
関数C5.0
学習データ(L0)を渡す
S40
関数C5.0
学習データ(L1)を渡す
S50

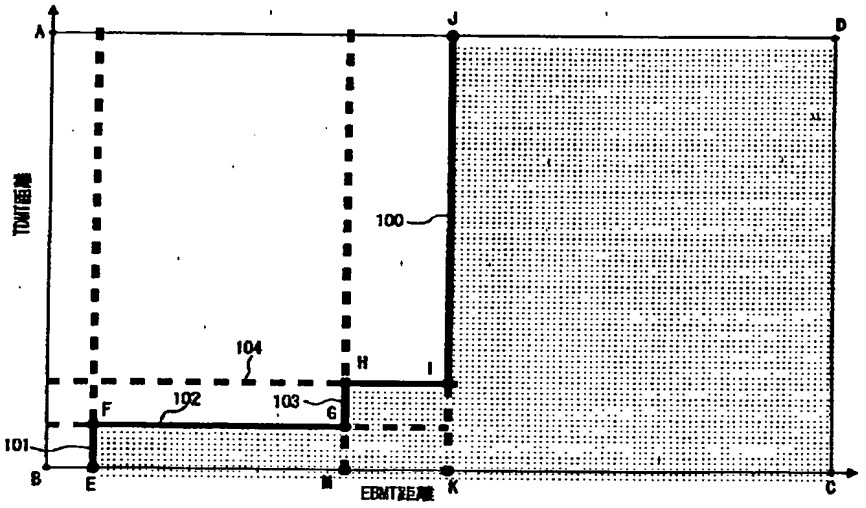
【図１０】


97
90
RAM
92
ROM
93
シリアルインタフェース
94
95
BS
CPU
91
CD-ROMドライブ
96
98
99

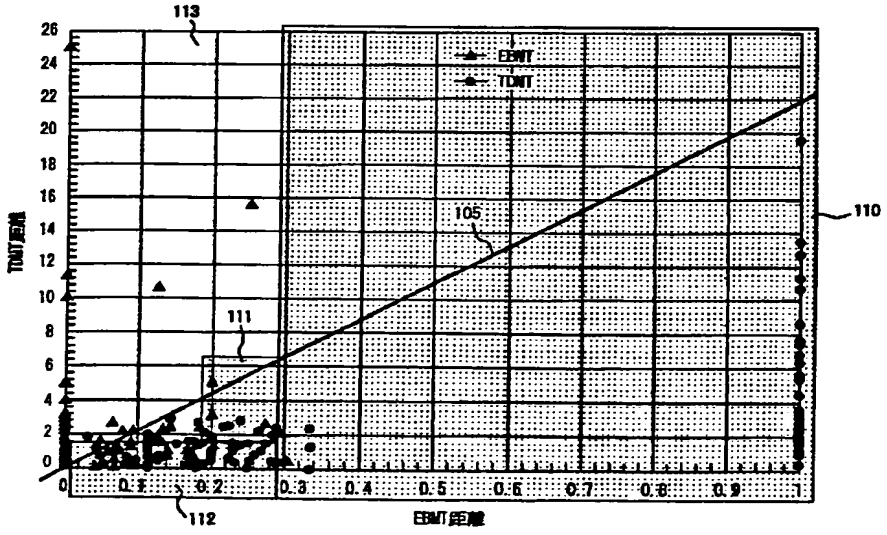
【図５】


L=(T, E, C)
T = {t1, t2, ・・・, tk, ・・・, tn}
E = {e1, e2, ・・・, ek, ・・・, en}
C = {c1, c2, ・・・, ck, ・・・, cn}
L0=(T0, E0, C0)
T0 = {t1, t2, ・・・, tk}
E0 = {e1, e2, ・・・, ek}
C0 = {c1, c2, ・・・, ck}
L1=(T1, E1, C1)
T1 = {tk+1, ・・・, tn}
E1 = {ek+1, ・・・, en}
C1 = {ck+1, ・・・, cn}

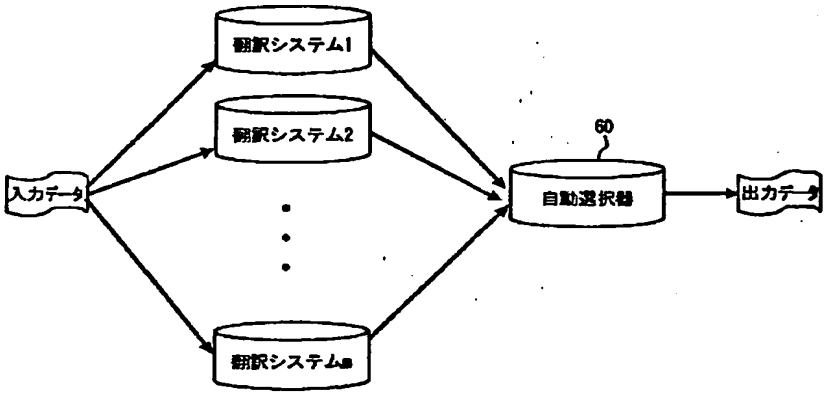
【図６】


A
J
D
TDMT距離
100
104
H
I
F
102
103
G
101
B
E
M
K
C
EBMT距離

【図７】


113
EBMT
TDMT
110
105
111
TDMT距離
112
EBMT距離

【図９】



フロントページの続き

(72)発明者　竹澤　寿幸
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

(72)発明者　山本　誠一
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

Ｆターム(参考)　5B091 BA11 CD13 CD15

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.